SAFESPRING SPECIFIKATION

SUNET Managed Cloud - Teknisk specifikation

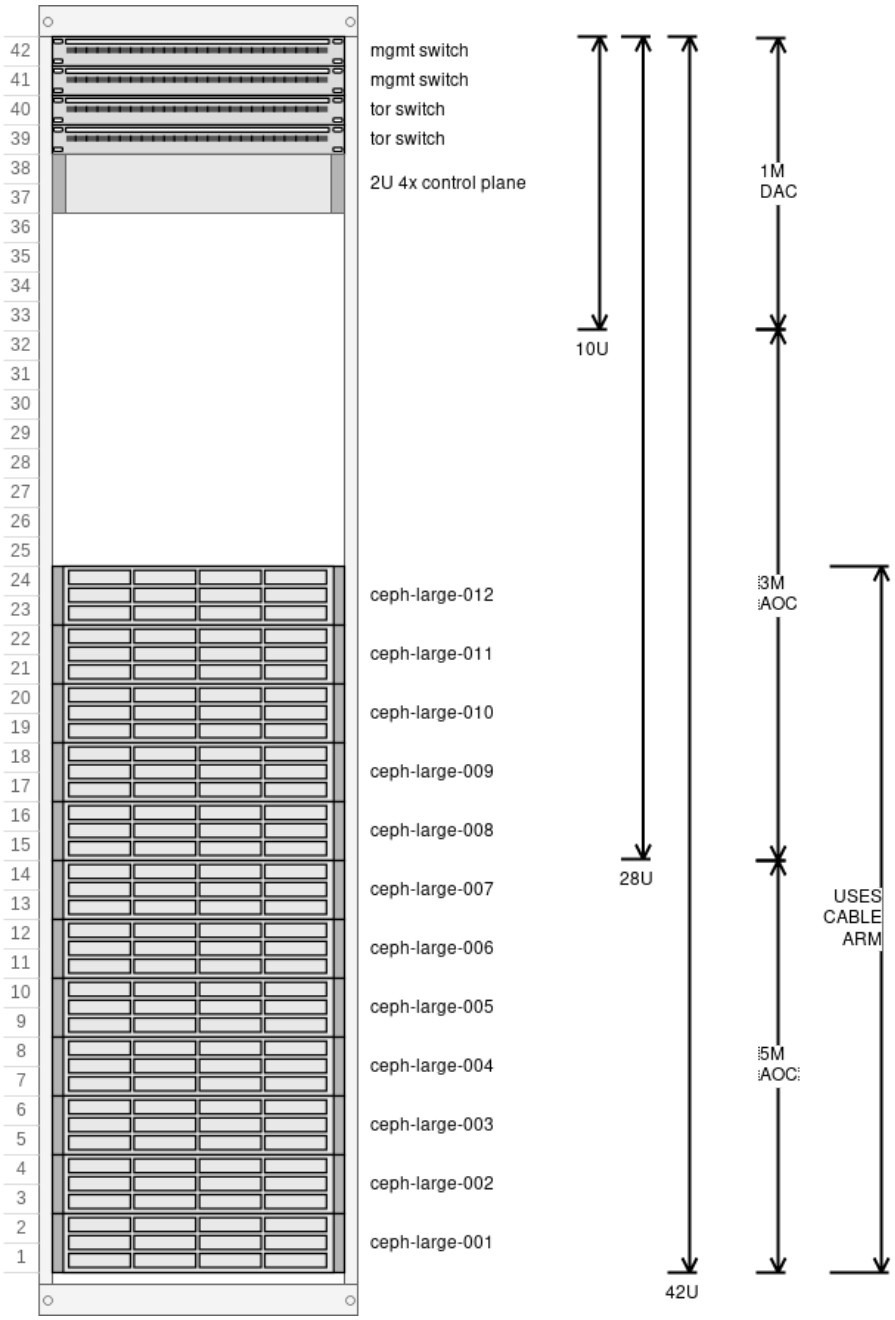
SUNET Managed Cloud är en tjänst från SUNET som ger möjlighet till storskalig lagring och resurser för virtuella maskiner i instansens (lärosätets/institutionens) eget datacenter. Lösningen är byggd på hårdvara som köps in av instansen och som installeras i instansens egen driftmiljö. På hårdvaran installerar sedan SUNET mjukvara som tillhandahåller lagring och resurser för kunden av använda. SUNET ansvarar för driften av hela stacken och kunden får tillgång till storskalig, objektbaserad lagring i sin egen miljö.

Blue Safespring AB
Smidesvägen 12
171 41 Solna

Tel: +46 (0)8-55 10 73 70
info@safespring.com
Number of pages: 11

Innehåll

Design


42
41
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
mgmt switch
mgmt switch
tor switch
tor switch
2U 4x control plane
ceph-large-012
ceph-large-011
ceph-large-010
ceph-large-009
ceph-large-008
ceph-large-007
ceph-large-006
ceph-large-005
ceph-large-004
ceph-large-003
ceph-large-002
ceph-large-001
10U
28U
42U
1M
DAC
3M
AOC
5M
AOC
USES
CABLE
ARM

Infrastruktur

	Managed Storage	Managed Compute	Kontrollnoder
CPU cores			
Memory			
Storage			
Connectivity	6 x 100GbE/4 x 25GbE		
Raw capacity	4 PB		
Capacity with Erasure Coding (4+2)			
Capacity with 3x replication			
Upload/Download speed	100 Mbits	-	-
Rack units	24 U	6U	
Power Consumption			
Access protocols	Ceph Block Storage, S3		

Nätverk

Management switches

Product	Product Description	Pcs
SSE-G3648BR-001	Layer 2/3 1/10G Ethernet SuperSwitch (länk till produktblad)	2
PWS-FRU-050	Spare Power Supply for SSE-G3648BR - reverse airflow, HF, RoHS	2
SFT-CLSPL1G-002	Cumulus-Linux SW 1G perpetual license with 3 yr Cumulus SnS	2

ToR switches

Product	Product Description	Pcs
DCS-7280SR2-48YC6-R	Arista 7280R2, 48 25GbE SFP and 6 x 100GbE QSFP switch, rear to front air, 2 x AC and 2 x C13-C14 cords	2
LIC-FIX-2-FLX-L	FLX-Lite L3 License for Arista Fixed switches, 40-132 port 10G OSPF, ISIS, BGP, PIM, Up to 256K Routes, EVPN, VXLAN	2

Cabling

Product	Product Description	Use	Pcs
CAB-Q-Q-100G-1M	100GbE QSFP to QSFP twinax copper cable, 1M	TOR1-TOR2	2
CAB-S-S-25G-1M	25GbE SFP25 to SFP25 twinax copper cable, 1M (works at 10G)	TOR-MGMT_SW	4
CAB-S-S-25G-1M	25GbE SFP25 to SFP25 twinax copper cable, 1M	TOR-COMPUTE	8
AOC-S-S-25G-3M	25GbE SFP25 to SFP25 Active Optical Cable, 3m	TOR-STORAGE	10
AOC-S-S-25G-5M	25GbE SFP25 to SFP25 Active Optical Cable, 5m	TOR-STORAGE	14
CAT5E-1M	1M CAT5E Ethernet cable	MGMT-COMPUTE	8
CAT5E-3M	3M CAT5E Ethernet cable	MGMT-STORAGE	10
CAT5E-5M	5M CAT5 Ethernet cable	MGMT-STORAGE	14

Arista - 25G Optics and Cables: Q&A Document
https://www.arista.com/assets/data/pdf/Arista25G_TC_QA.pdf

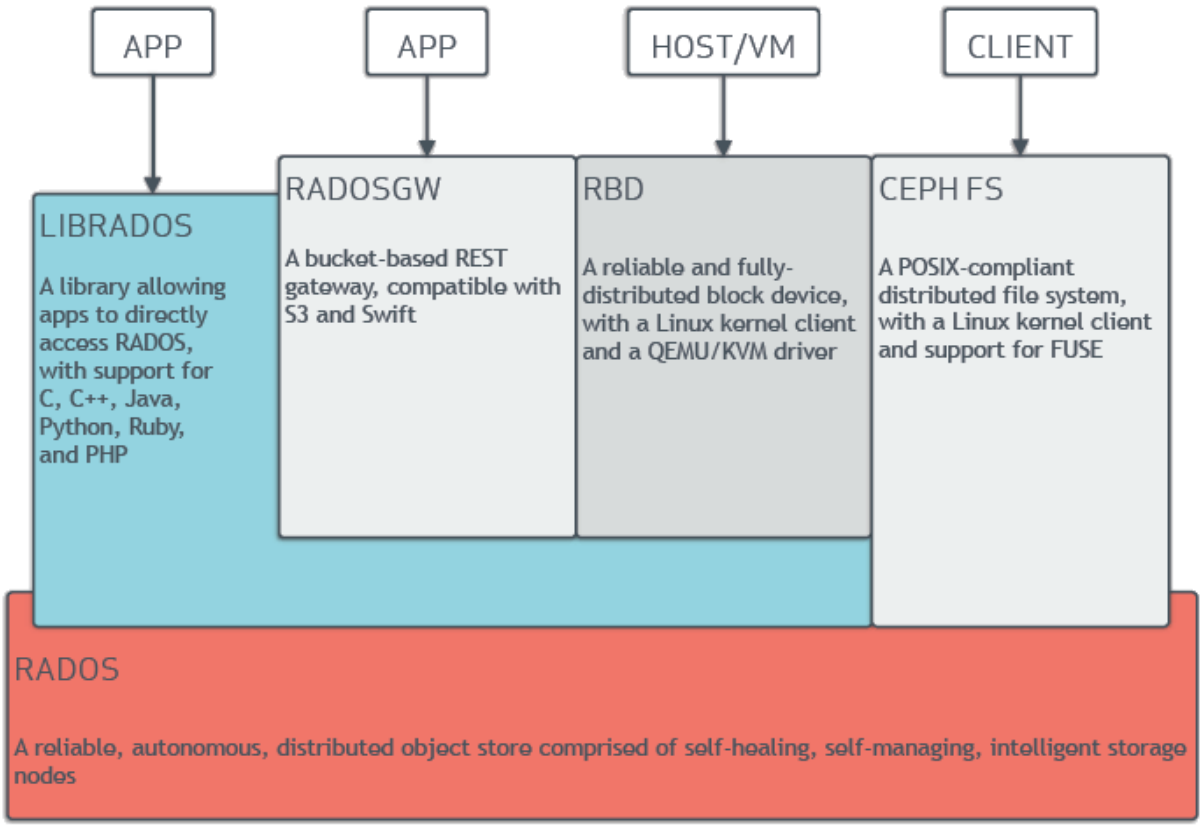
Mjukvara

Managed Storage - Ceph

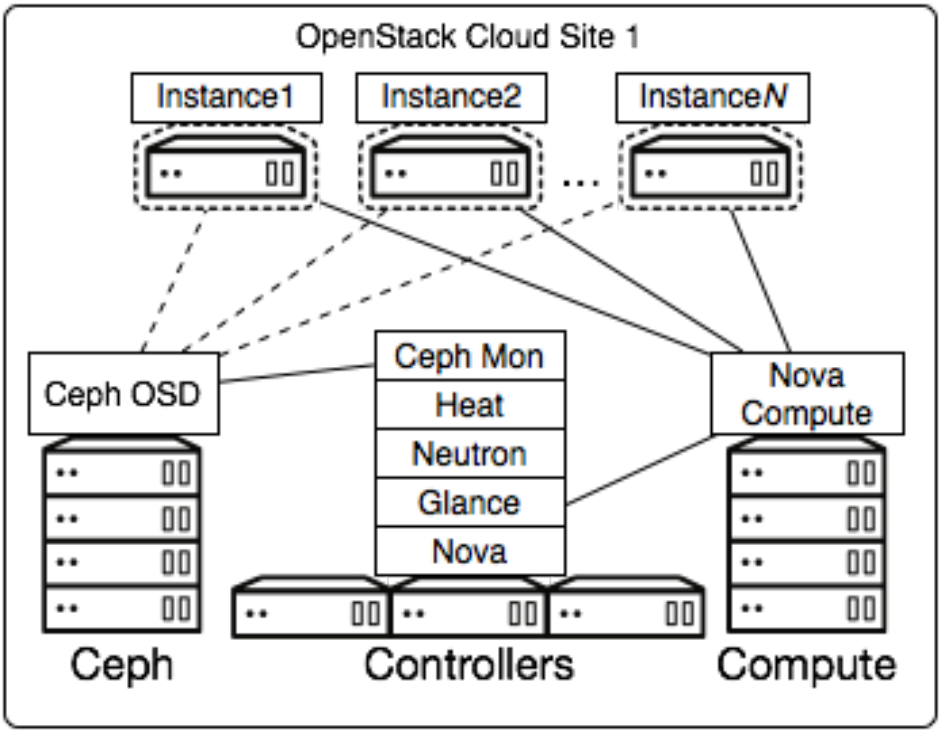
Det centrala i Ceph är CRUSH-algoritmen som är en hash-algoritm som möjliggör för anslutande klienter att ansluta direkt till storage noderna utan behov av någon administrativ nod emellan. Vilken nod som det önskade datat räknas ut med CRUSH-algoritmen och om den primära noden är nere så visar CRUSH vilka noder där datat också finns. På så vis undviks SPOF (single point of failure) och flaskhalsar i administrationsnoder och gör också att lösningen skalar väldigt bra.

Ceph består av ett antal komponenter. Förutom själva storage noderna (OSD) så finns det monitor-noder (och radosgw-noder som ger tillgång till lösningen över S3-protokollet.

Bilden nedan visar över vilka gränssnitt det går att kommunicera med Ceph.



Managed Compute - OpenStack



Tjänster

Managed Storage

Managed Storage är en storskalig lagringslösning baserad på det välkända öppen-källkod-projektet Ceph. Versionen som används är Luminous. Gränssnittet till lösningen är av två varianter beroende på tillämpning: blocklagring eller objektlagring. Blocklagring använder Ceph Block Storage-protokollet medan objektlagringen är kompatibelt med S3.

Managed Storage kommer i tre varianter, beroende på vilka behov man har: **arkiv**, **large** och **fast**. Det går att leverera samtliga tre typer samtidigt från samma lösning.

Nyckelfunktioner:

1. **Byggd på marknadsledande Ceph-molnteknologi**
    a. Välj mellan olika lagringstjänster; Fast, Large eller Archive.
    b. Hög nätverksprestanda samt mjukvarudefinierade virtuella nätverk för att enkelt särskilja trafik

2. **Hög säkerhet**
    a. Rollbaserad accesskontroll till tjänsten. Möjlighet att integrera med identitetssystem
    b. Integrerat med SWAMID och FEIDE för utbildningssektorn
    c. Data krypteras både under transport (TLS) samt i datacenter på krypterade hårddiskar

Arkiv

Varianten är optimerad på pris, med möjlighet att lagra stora mängder data utan skenande kostnader. Lösningen är av typen objektlagring med S3-gränssnitt vilket är idag har blivit en de facto standard för interaktion med objektlagringslösningar. Protokollet använder ett REST API för att lägga till, ladda ned och ta bort filer i lösningen. Det finns många klienter och programvaror som använder protokollet tillsammans med en uppsjö bibliotek som möjliggör integration i programvara man utvecklar själv. Objektlagring gör att förbrukningen av tjänsten mäts i bytes som det totala antalet filer som ligger i tjänsten förbrukar.

Idag levererar Safespring en delad tjänst av typen Managed Storage Arkiv från SUNETs datacenter Stockholm B som går att bruka för kunder som är anslutna till SUNET.

Large

Lösningen är en block storage lösning vilket betyder att den presenterar volymer till andra servrar (virtuella eller fysiska) över kundens interna nätverk. Large använder roterande mekaniska diskar som lagringsmedia vilket gör lagringskostnaden lägre än varianten Fast. Volymer skapade i Large lämpar sig för mindre krävande applikationer eller applikationer med stora mängder data. Förbrukningen mäts i antal bytes som det totala antalet volymer som man har allokerat i tjänsten förbrukar.

Fast

Även den här varianten är av typen block storage, men med skillnaden att SSD-diskar används som lagringsmedia vilket ger hög hastighet för åtkomst. Volymer av den här typen lämpar sig bäst för operativsystemsvolymer eller mer krävande applikationer. Förbrukningen mäts i antal bytes som det totala antalet volymer som man har allokerat i tjänsten förbrukar.

S3 API (Ceph Luminous)
Varianter arkiv, large, fast
Vi har varianten arkiv per nå (se docs till SUNET).

Managed Compute

Managed Compute använder sig av OpenStack för att tillhandahålla möjligheten att köra virtuella maskiner i plattformen. OpenStack kan antingen manageras genom ett grafiskt gränssnitt i en browser eller över ett dokumenterat API.

Plattformen tillhandahåller färdiga avbildningar av de vanligaste operativsystemen. Resurser till de virtuella maskiner man startar använder flavors, eller fördefinierade resursprofiler som anger hur många vCPU:er och hur mycket minne som maskinen skall få. Det finns också möjlighet att binda flavors till specifika resurser i plattformen, som t ex GPU-bestyckade noder, för att kunna få högre prestanda för vissa krävande applikationer som t ex deep learning. Notera att GPU-bestyckade noder behöver köpas separat.

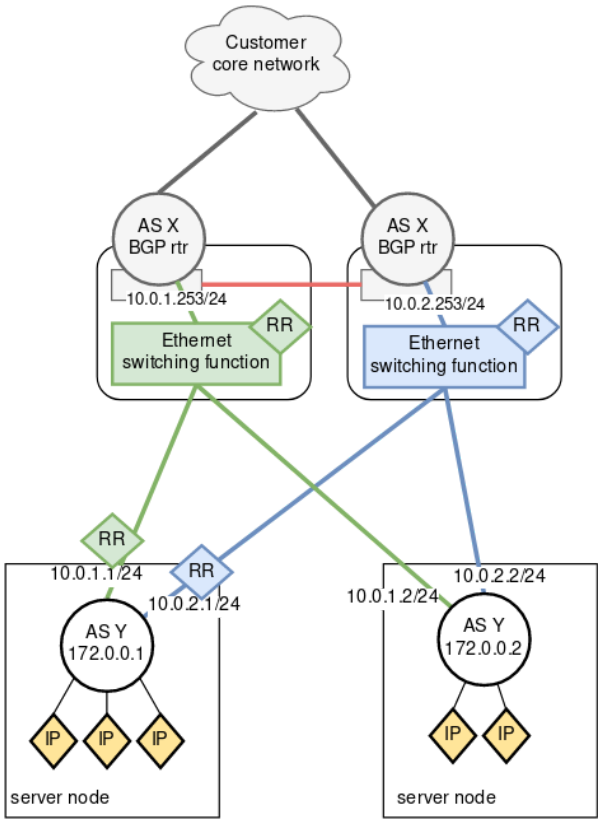
Nätverkslagret realiseras med projektet Calico vilket är en BGP-baserad nätverkslösning för storskalig hantering av nätverksresurser till de virtuella maskinerna. Maskinerna kan kopplas till ett publikt nätverk (som ger en publik IP-adress och nåbarhet över internet), ett helt privat nätverk (som ger en privat adress utan möjlighet att nå internet) eller ett privat nätverk med inbyggd NAT-funktion som ger den virtuella maskinen möjlighet att nå ut på internet - även om den inte kan nås utifrån.

Nätverksdesign

SUNET sköter TOR-switcharna där anslutningen till SUNETs nätverk sker till lösningen. Safespring får läsrättigheter för SNMP i den utrustningen för att snabbt kunna utesluta fel i anslutningen vid felsökning. Ändringar i utrustningen initieras av båda parter men utförs av SUNET.

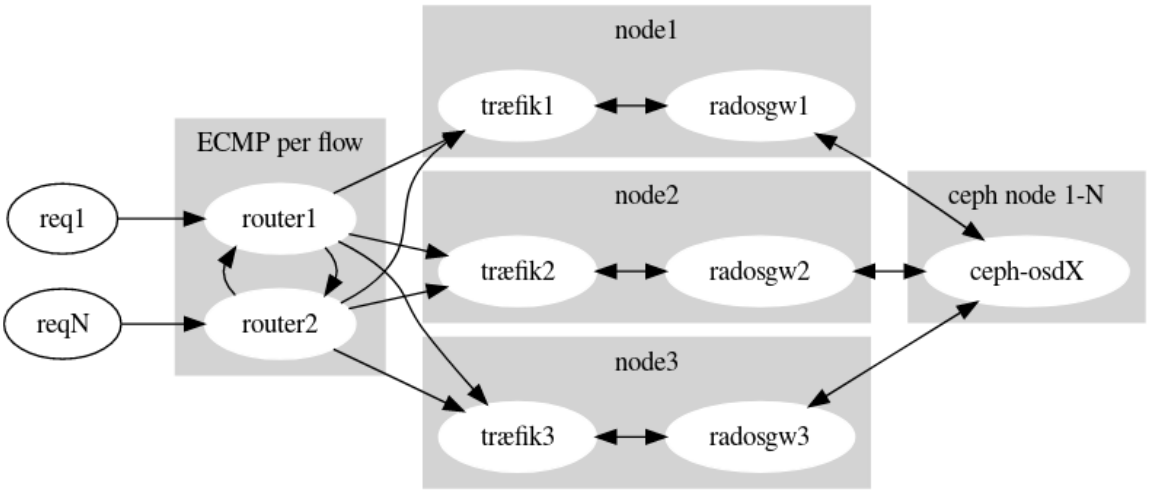
Safespring sköter management-switcharna. SUNET har läsrättigheter för SNMP till de switcharna för enklare felsökning.

Nedan är en konceptuell bild som visar BGP-routrar och konnektivitet till lösningen som placeras i kundens driftmiljö. Det är en relativt enkel L2-lösning med multipla distinkta L2-kanaler utan användande av L2-redundanta protokoll. En BGP router på varje servernod ger redundans genom två separata vägar till de andra servernoderna. Det finns också två vägar till kundens nätverk genom TOR edge routrar ut ur lösningen.
Två routereflektorer per L2-switchplan delar routinginformation genom iBGP. På så vis undviks behovet av en full mesh-lösning.



Redundansdesign

För att säkerställa tillgänglighet och effektiv överföringshastighet är lösningen uppsatt som följer med lastbalanserare (Træfik) till de bakomliggande Ceph-noderna. När ett anrop kommer inte till lösningen så lastbalanseras anropen på flow-nivå. Genom att sätta upp lastbalanserare och RadosGW-noder i strikta par på samma hårdvarunod så ökas prestandan vid överföring eftersom att trafiken behöver ta färre hopp över nätverket. Om en RadosGW-nod skulle gå nod så kan den framförliggande lastbalanseraren skicka trafiken till en annan hårdvarunod - men i normalfallet så behövs inte det varför prestandan ökas.



Redundansnivåer i själva lagringsklustret bestäms av hur Ceph-klustret är konfigurerat. Standardinställningen är att använda Erasure-kodning med 4 datablock och 2 paritetsblock (RAID6) vilket ger en hög tillgänglighet med bara en 50% ökning av utrymme på de bakomvarande diskarna. Det är också möjligt att sätta upp lagrinspooler med andra inställningar, såsom 3 replikat av varje block vilket ger hög säkerhet och högre åtkomsthastigheter till priset av ett högre användande av lagringsyta (200%).